

DOCUMENT RESUME

ED 409 360

TM 026 892

AUTHOR Nekovei, Deanna L.
 TITLE Assessment of Beginner Teacher Performance Utilizing Video Technology.
 PUB DATE 24 Mar 97
 NOTE 30p.
 PUB TYPE Reports - Evaluative (142)
 EDRS PRICE EDRS Price MF01 Plus Postage. PC Not Available from EDRS.
 DESCRIPTORS *Beginning Teachers; Cost Effectiveness; Educational Technology; Elementary Education; *Elementary School Teachers; Holistic Evaluation; *Performance Based Assessment; Performance Factors; *Scoring; *Teacher Evaluation; *Videotape Recordings

ABSTRACT

A study was conducted to investigate the feasibility of using video as a method of beginning teacher observation. In addition, the study investigated whether or not trained observers (raters) could score beginner teacher performance videotapes reliably using a 4-point holistic scoring scale. To this end, the researcher collected 24 beginning teacher performance videotapes and had 12 elementary school teachers, the raters, score each videotape using a four-point holistic scoring scale that was developed for the study. In sum, it appears that videotape performance assessment that uses a holistic scoring scale is a viable and cost effective method of teacher evaluation. It was also learned that camera condition is not a significant factor in terms of obtaining consistent scores on the performance videotapes. A 20-to-30 minute tape appears to provide ample teacher performance information for obtaining consistent judgments by raters. (Contains 2 figures, 13 tables, and 32 references.) (Author/SLD)

 * Reproductions supplied by EDRS are the best that can be made *
 * from the original document. *

Assessment of Beginner Teacher Performance Utilizing Video Technology

Deanna L. Nekovei

Texas A&M University-Kingsville

Kingsville, Texas 78363

March 24, 1997

PERMISSION TO REPRODUCE AND
DISSEMINATE THIS MATERIAL
IN OTHER THAN PAPER COPY
HAS BEEN GRANTED BY

Deanna L. Nekovei

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

☒ This document has been reproduced as
received from the person or organization
originating it

☐ Minor changes have been made to
improve reproduction quality

• Points of view or opinions stated in this
document do not necessarily represent
official OERI position or policy

BEST COPY AVAILABLE

ABSTRACT - The purpose of this study was to investigate the feasibility of using video as a method of beginner teacher observation. Additionally, this study investigated whether or not trained observers (raters) could reliably score beginner teacher performance videotapes using a four point holistic scoring scale. To this end the researcher collected 24 beginner teacher performance videotapes and had twelve elementary school teachers, the raters, score each videotape utilizing a four point holistic scoring scale that was developed for the purposes of this study. In sum, it appears that videotape performance assessment that utilizes a holistic scoring scale is a viable and cost effective method of teacher evaluation. Lastly, it was learned that camera condition was not a significant factor in terms of obtaining consistent scores on the performance videotapes.

1 INTRODUCTION

1.1 Background

It has recently become apparent that teacher licensure tests that rely on the traditional paper-pencil, multiple choice format of evaluation are leaving much out of what it means to be a good teacher. Paper and pencil exams are able to capture reading comprehension; mathematic ability; knowledge of pedagogy; content area knowledge; and knowledge of learners all of which have content validity. However, they are not able to capture how teachers actually apply this knowledge in a classroom (Haertal, 1990). It is here that performance assessment can have its greatest impact. But, what is performance assessment? Berk (p. ix, 1986) states that "performance assessment is the process of gathering data by systematic observation for making decisions about an individual." Further, Evertson and Holley (p. 90, 1981) note that "classroom observation gives us a view of the climate, rapport, interaction, and functioning of the classroom available from no other source."

National Evaluation Systems (NES) is taking a leading role in the development of performance assessment measures for instance in order to receive a permanent teaching certification in New York State, teachers

will be asked to submit a videotape of a class that he or she has taught with a completed form explaining the background, goals, and context for the lesson and providing brief comments on the success of the session (*The State Education Department, University of the State of New York, p. 2, 1991*).

This example briefly illustrates the role that performance assessment measures are taking and alludes to the potential value this method may have in the evaluation of teaching ability, especially hands-on applications of teaching techniques. Thus, it should not be a question of whether or not teachers will be evaluated but how they will be evaluated given the fact that

We make judgments all the time, judgments about ourselves and what we do and about others and what they do. And we, in turn, are being judged by others. We cannot escape evaluation. Every choice, every decision—to speak or not, to use this example or that— involves an evaluation, automatic or deliberate (*Millman, p.12, 1981*).

Further, teacher evaluation traditionally has been conducted by means of direct observation using an analytic scoring format. However, this approach to assessing teacher

competency neglects many aspects of teacher behavior and classroom learning such as teacher warmth and creativity in lesson implementation.

Additionally, Teacher licensure/certification personnel should consider relevant research on teaching and the practical experience of qualified experts (e.g., teachers, faculty, administrators) in the construction of beginning teacher assessments. Thus, they should be well versed in the following research areas: teacher effectiveness, expert versus novice teaching, and the process of learning to teach.

1.2 Experts versus Novices

"What makes expert problem solvers better than novices? Clearly, experts in a domain have had more training and practice than have novices, but what exactly is it that experts have learned? Have they learned to reason better in general, or perhaps become more skilled in applying heuristic search methods, such as means-ends analysis" (Holyoke, 1990, p. 128)?

De Groot (1965) in pioneer research in expert vs. novice chess players found that expert chess players did not have better reasoning ability, nor did they perform better on means-ends analysis. It seemed that expert players instead considered fewer, not more moves than novice players. In addition, experts used past knowledge to locate good moves in a short space of time, and used the greatest amount of time contemplating good moves, not considering all possible moves. Additionally, Chase and Simon (1973) discovered that expert chess players seemed to store knowledge of past chess board configurations in large meaningful units called chunks. Chunking enabled experts to encode information in a timely and accurate fashion only when the chess pieces were arranged in a plausible board configuration, not randomly positioned.

Thus, experts have learned to categorize information in a more abstract form and in larger chunks in order to limit search time. It also seems that research in the area of experienced and beginner teachers also follow this pattern.

1.3 Preactive Tasks

Teachers make plans that encompass relatively short periods of time (a single lesson, daily, weekly, and unit planning) and others which cover more lengthy periods of time (quarterly, semesterly, and yearly). However, there is research that suggests that experienced and beginner teachers plans may vary significantly. Experienced teachers tend to view lesson planning in a nonlinear fashion whereas, beginner teachers see lesson

planning in a linear fashion (Clark and Peterson, 1986). It seems beginner teacher lesson planning requires more time and effort than experienced teachers planning. Borko and Livingston (1989) state that experienced teachers are able to "plan more quickly and efficiently than novices because they are able to combine information from existing schemata to fit the particulars of a given lesson" (p.490). Additionally, beginner teacher planning lacks contingency plans, relies on concrete methods and activities which is unlike experienced teachers (Borko, Livingston, McCaleb and Mauro, 1988 and Housner and Griffey, 1985).

Competent teachers are able to develop lessons that build on past knowledge. They also tend to call attention to similarities and the sequences between old and new information, and across subject matter (Biehler and Snowman, 1986). Successful lessons meet students educational and developmental needs and interests (Brophy and Good, 1986). Competent teachers also set high expectations for student achievement and behavior, but at a level that each student is capable of meeting with a little effort (Hunter, 1982). It seems beginner teachers have trouble with defining appropriate expectations for children.

Experienced teachers are able to tailor learning materials to meet the individual needs of the students, but it has been found that beginner teachers often have trouble in this area (Schram, Feiman-Nemser, and Ball, 1989). Beginner teachers also have difficulty relating the "big picture" of a lesson, topic, theme to students. It seems beginner teachers often do not consider students prior knowledge and performance on subject matter unlike experienced teachers (Shulman, 1989). This is probably the result of beginner teachers lack of experience in grades other than the one he/she is currently teaching. Additionally, few teacher colleges provide direct instruction in the scope and sequence of curriculum across grades and subject matter. College educators seem to feel that pre-service teachers will pick up this information in their field experiences. Needels (1991) also notes that experienced teachers are able to draw on student knowledge while inexperienced teachers tend to focus on learning that is acquired in schools.

1.4 Interactive Tasks

Teachers are actively involved in the teaching process in this phase. Teacher thinking often involves thoughts about pupil characteristics, lesson objectives, content organization, and the classroom environment (Clark and Peterson, 1986). Again, the research in this area is compatible with the beginner and experienced teacher differences such that experienced teachers are able to sift out the important from the trivial information and to chunk this information in a relatively short space of time in order to make quick

decisions.

Competent teachers are able to create a learning environment that minimizes student misbehavior through positive reinforcement, modeling, and holding students accountable for their actions, and making sure students understand the lesson (Biehler and Snowman, 1986; Hunter, 1982; Evertson, Emmer, Clements, Sanford, Worsham, 1989; and Duke and Meckel, 1984). Competent teachers establish clear rules and consequences for breaking rules. They also set up the physical environment in a way that is conducive to learning such as arranging class seating in such a way that students are visible to the teacher at all times. They also group students according to appropriate social and learning needs. Additionally competent teachers are often characterized as having "eyes in the backs of their heads" which is called "withitness" (Kounin, 1970) in the teaching literature. They tend to use "I" messages as opposed to "You" messages (Ginott, 1971 and Gordon, 1974). For instance, an I message would be "I don't like it when you forget to put away your things." Whereas, a you message would be "You also make such a mess..." They use face saving strategies such as privately conferencing with misbehaving students and making sure the student is aware of their actions, and why their actions are inappropriate along with a discussion of ways to solve the problem. Lastly, competent teachers strive to maximize on task learning by minimizing the need and time it takes for transitions.

In general, competent teachers ignore minor problems and promptly attend to serious problems by using signals such as eye contact, head shaking, making use of proximity, avoiding threats, and repeating oneself. It is this area (i.e., classroom management) that beginner teachers often have the most difficulty, especially in dealing with misbehavior and establishing respectful learning/teaching relationships with students. That is beginner teachers have trouble establishing themselves as a warm, caring professional, not another playmate. This is substantiated by Needels (1991) who suggests that inexperienced teachers focus more on affective components of teacher-student interactions while experienced teachers emphasize teacher feedback to students.

Thus, it seems that classroom management occupies a greater amount of time and attention for beginner teachers. Needels (1991) notes that student teachers and beginner teachers, who viewed a thirty minute video of a first grade lesson seem to place a greater emphasis on classroom management issues as opposed to experienced teachers. Experienced teachers instead focused on lesson content and strategies used by the first grade teacher in the video.

Additionally, Sabers, Cushing, and Berliner (1991) suggest that when novice teacher

and beginner teachers view a video of a classroom lesson they have difficulty differentiating between important and unimportant stimuli as opposed to experienced teachers, who were able to make meaning out of the video lesson. It seems that inexperienced teachers are unable to tell what makes a good lesson and tend to place an inordinate amount of emphasis on instructional materials. However, experts are able to "a) monitor and comprehend the events presented, b) interpret the instructional strategies used, c) hypothesize reasons for behavior seen, and d) offer solution strategies for problems identified" (p. 84).

Competent teachers are able to make subject matter understandable and memorable for students because they usually know the subject matter better than beginner teachers, and are able to approach it from a pedagogical framework (Shulman, 1987). That is they are able to pin point why students make the errors they do, what leads to or causes some of their misunderstandings and to develop ways to teach lessons that avoid these pitfalls. Student teachers (Haymore, 1987) and beginner teachers (Grossman and Richert, 1988) are just beginning to develop these skills. However, experienced teachers show a greater understanding of learning and its interconnectedness with other events (Needels, 1991).

Further, competent teachers evaluate readiness to learn a new skill, concept, problem by asking questions, reviewing past material, testing students knowledge of material, and reteaching as necessary (Biehler and Snowman, 1986). When students are ready to move on, competent teachers consider past knowledge and abilities of the learner in their preparation of lessons. Additionally, competent teachers consistently model learning, guide students through a series of problems (situations, concepts), provide practice, and are prompt with feedback (Hunter, 1982). They offer metaphors and analogies to facilitate transfer to new learning situations (Siegler, 1991).

However, beginner teachers may not be so clear with their expectations about what is to be learned. They frequently change what they do in class and the ways they do them such as lecturing and presentations (Leinhardt and Greeno, 1986). This may be the result of beginner teachers grappling with the subject matter at hand along with ways to make it meaningful to students. In contrast, experienced teachers show a more consistent pattern of behavior in their instruction.

1.5 Postactive Tasks

Competent teachers reflect on whether or not they are successful in meeting student needs, during as well as after a lesson (unit). They tend to use multiple forms of evaluations such as questioning students for understanding, providing practice and review of material and they make use of formative and summative exams in order to tell if

students are meeting teacher expectations of learning (Hunter, 1982 and Biehler and Snowman, 1986).

However, beginner teachers have difficulty reflecting on their teaching and often mention incidental things such as did they make good use of the blackboard (Borko and Livingston, 1989). As mentioned earlier they tend to be overly concerned about classroom management issues (Needels, 1991). However, experienced teachers place more emphasis on student understanding and place less emphasis on classroom management (Borko and Livingston, 1989). Additionally, experienced teachers make use of more "if-then" thinking (Berliner and Carter, 1989, p.56). Experienced teachers often reflect on student-teacher interactions such as did the teacher ask probing questions, make use of wait time, show acceptance of student answers (Hunter, 1982), make use of familiar vocabulary, and provide clarifications, explanations of unfamiliar words and terms (Needels, 1991). In general, beginning teachers mention far more inconsequential events than do experienced teachers (Needels, 1991). This may be the result of the fact that for the beginning teacher everything is new and therefore cause for discussion.

1.6 Conclusion

In closing, beginner teachers are slowly developing the skills necessary to teach successfully, and while they may not be as proficient as experienced teachers "we [still] should expect beginning teachers to be able to: plan lessons that enable students to relate new learning to prior understanding and experiences; develop rapport and personal interactions with students; establish and maintain rules and routines that are fair and appropriate to students; arrange the physical and social conditions in the classroom in ways that are conducive to learning and that fit the academic task; represent and present subject matter in ways that enable students to relate new learning to prior understanding and that help students develop metacognitive strategies; assess student learning using a variety of measurement tools and adapt instruction according to the results; and reflect on their own actions and students' responses in order to improve their teaching" (Reynolds, 1992, p.26).

1.7 Purpose of the Study

It is perceived that video could potentially be a more cost effective method of teacher evaluation as opposed to direct observation. Thus, the purpose of this study was to investigate the feasibility of using video as a method of beginner teacher observation. Additionally, this study investigated whether or not trained observers could reliably

score beginner teacher performance videotapes using a four point holistic scoring scale.

1.8 Limitations

This study was designed to evaluate beginner teacher performance videotapes utilizing a four point holistic scoring scale. For this reason, the results may or may not be applicable to experienced teachers.

This study also was not designed to compare direct versus indirect methods of teacher evaluation. Further, the holistic scoring method of evaluation does not allow for comparisons between people outside of the group evaluated, since holistic scoring is essentially a method of ranking.

In reference to videotaping, this study sought to determine the most cost effective method of videotaping beginner teacher performance. While at the same time minimizing loss of information due to not having the beginner teacher and/or students in view. Thus, this study positioned two wide-lens cameras (without operators) in the room, one camera at the front of the room and one camera at the back of the room. This produced three videotaping conditions: video camera at the front of the room, video camera at the back of the room, and the two video cameras running simultaneously (front and back of room) in order to produce the dual image condition. This approach, however, does not permit information to be obtained by close-up shots such as zooming in on something written on the blackboard. Nor does it allow for tracking a beginner teacher's movements.

Lastly, this study was not designed to compare direct observation versus videotaped observation.

2 METHODOLOGY

2.1 Introduction

The goals of this study were two-fold, first to assess the feasibility of using videotapes to evaluate beginner teacher performance, and secondly, to determine if a reliable holistic scoring scale could be developed to evaluate the videotaped observations. Thus, this section will be broken down into two main parts or phases. Phase one will discuss the collection of teacher observation videotapes. Whereas, phase two will focus on the development of a holistic scoring scale that was used to evaluate the teacher observation videotapes.

2.2 Phase One

2.2.1 Participants

The participants were twenty-four (22 females and 2 males) undergraduate education majors attending the University of Massachusetts at Amherst. The participants were in the process of completing student teaching in the public schools within the vicinity of Amherst.

Permission to videotape student teachers in the public schools was obtained from five districts, eleven schools agreed to participate. Informed consent forms were obtained for each child in the student teacher's class. Four types of consent forms were used that of, blocked, regular, Spanish, and Chinese. The blocked version stipulated that the researcher would block out any child whose parents did not want to be in the videotape, but who did want the child to participate in the lesson. The regular version only requested that the child be allowed to be videotaped during a class lesson. The last two versions were essentially the same as the regular version except they were written in Spanish and Chinese, respectively. Additionally, each student teacher submitted a draft of their lesson plan, and the consent of each student teacher was obtained prior to videotaping.

Lastly, the participants were informed of their right to withdraw from the study at any time. Throughout the rest of this study the participants will be called beginner teachers, since they were just completing their teaching degrees at the time of videotaping.

2.2.2 Procedures

The participants were videotaped using two cameras, one at the front of the room and one at the back of the room. Additionally, the two videotapes were placed side-by-side to produce a third condition, the dual image condition. The third condition was essentially three videotapes given that this condition used two cameras. Thus, there were three videotaping conditions: video camera placed at the front of the room, video camera placed at the back of the room, two video cameras used simultaneously (front and back of room) in order to produce the dual image condition.

This study was interested in trying to determine the most cost effective method of videotaping beginner teacher performance. Thus, the cameras were mounted on tripods and positioned at the front and back of the room in order to obtain the most information possible on beginner teacher performance utilizing a wide lens. This enabled the researcher to let the cameras run independently of a camera operator while at the same

time still being able to capture classroom information. This design had the advantage of obtaining the maximum amount of information possible about which type of camera position and format was best suited for accurately scoring beginner teachers' performance videotapes while minimizing the number of participants necessary to conduct the study.

The participants were videotaped by the researcher while conducting a typical classroom lesson, in other words, no attempt was made to manipulate the mode of instruction. Figure 2.2.2 illustrates general camera positions used in videotaping beginner teacher performance.

2.2.3 Phase Two

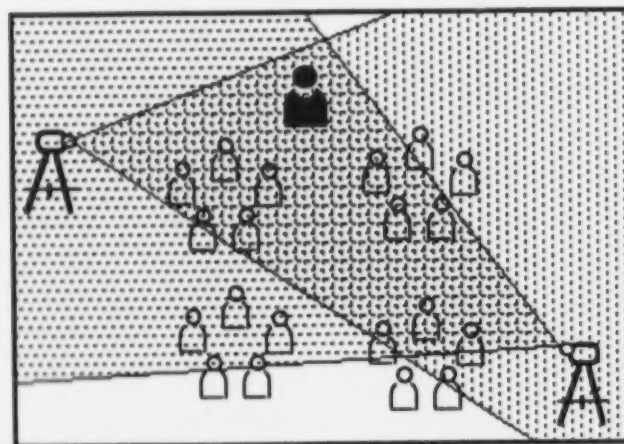
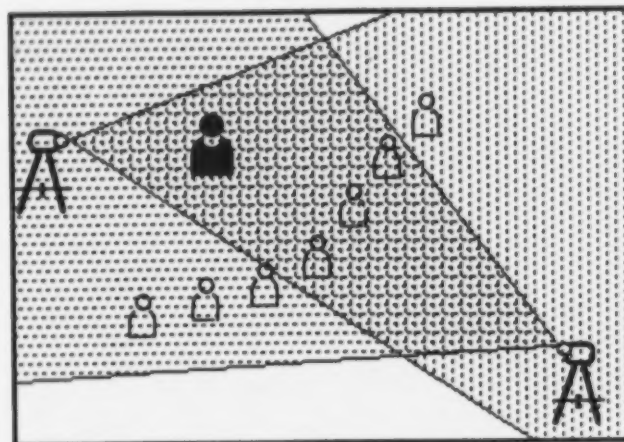
Introduction It was the goal of phase two to develop a holistic scoring scale and to train raters to rate (score) the beginner teacher performance videotapes that were collected in phase one of this study.

Participants A total of twelve female elementary school teachers from Western Massachusetts participated as raters (judges) for the purpose of this study. Each teacher rated nine beginner teacher performance videotapes holistically. The age of the participants ranged from 25 years to 52 years (mean=44 years). The number of years the teachers had taught at the elementary level ranged from 4 years to 26 years (mean=19 years). Further, nine of the 12 teachers taught in grades kindergarten through grade four. While the remaining three teachers taught in more than one grade (e.g., art teacher).

Lastly, for the remainder of this study, the elementary school teachers will be referred to as Raters given that this was the primary role that they fulfilled.

2.2.4 Procedures

This section is based on a holistic scoring manual developed by Eileen Peters (1979) of Educational Testing Service for the Massachusetts State Board of Education's Basic Skills Improvement Policy. This manual was developed to educate and train people to conduct holistic scoring sessions for the assessment of written text. Additionally, White (1985) and the Focused Holistic Scoring of TABS Writing Samples, Exit Level. Texas Assessment of Basic Skills (1980) that was produced by the Texas Education Agency provided valuable information pertaining to the organization of a holistic scoring session. It was the goal of this study to apply the knowledge of holistic scoring for written assessment to videotaped beginner teacher performance assessment. The following sections discuss in depth the holistic scoring process.



Legend:



Students



Teacher



Camera

Figure 1: Illustrations of camera positions for whole group (top) versus small group (bottom) instructions.

Scale Development The development of the holistic scoring scale that was used by the raters to judge the beginner teacher performance videotapes was comprised of four primary components, that of, the literature review, the lesson plans submitted by the beginner teachers, and the collected beginner teacher performance videotapes all of which will be described in this section.

The initial stages of scale development were mainly composed of reviewing and citing literature on effective teaching, what constitutes expert versus novice teaching, what is competent beginner teaching, and methods of teacher evaluation with and without the use of video. The review of the literature was essentially done in order to ascertain what level of performance can and should be expected from beginner teachers.

Lesson plan format and collected beginner teacher performance videotapes from phase one of the study were reviewed and detailed notes were taken on the use of whole group versus small group instruction, teaching techniques and strategies, and instances of good and bad teaching performance (e.g., transitions smooth or chaotic). The aforementioned areas of initial scale development were undertaken because research in holistic assessment of written text suggests that these steps are not only necessary, but crucial to the development of a valid and reliable scale (Texas Education Agency, 1980; White, 1985; and Williamson, 1993).

Following in the footsteps of holistic assessment for written text, the literature review was used in the preliminary stages of developing the score points and the objectives. Next, the objectives were revised and a draft description of the score points were developed based on the submitted lesson plans. Lastly, in light of the information gained from the collected beginner teacher performance videotapes, the researcher refined the description of score points that are in essence based on the objectives. Tables 1 and 2 illustrated the score points and the objectives, respectively. Note that there are four score points this was done so that raters would be forced away from choosing a middle score, and additionally to create greater statistical variability in the scale.

Selection of Training Videotapes Prior to the holistic scoring session, six beginner teacher performance videotapes were randomly selected, viewed, and scored utilizing the holistic scoring scale that was developed for this study. The videotapes selected were used to calibrate the raters, in other words, these videotapes acquainted the raters with the full range or types of videotapes that they would be scoring in order to reach a consensus on what videotapes constituted a score of four as opposed to a score of three, two, or one. They also, familiarized the raters on the types of problems that they may encounter such as degraded audio/visual quality. The raters were advised that problem

Table 2: Instructional objectives.

<i>Objectives</i>
<i>Instructional Orientation</i> Ties in past learning with new learning to build on students' knowledge and to ensure continuity and sequencing of learning. Gives clear, concise, reasonable directions to students. Communicates expectations for successful performance to students. Lesson follows a logical pattern.
<i>Instructional Method/Delivery</i> Subject matter appropriate for grade and ability level of students. Groups students in an appropriate manner (individual, small, and whole groups). Uses effective instructional approach (lecturing, modeling, questioning, experimentation, role play). When appropriate uses multisensory approaches (tactile, visual, auditory, kinetic). Communicates effectively with students by making use of verbal and nonverbal cues (tone of voice, facial expression, gestures, movement about the room). Stimulates and encourages creative, critical thinking, and problem solving skills. Give constructive feedback frequently and promptly. Demonstrates enthusiasm.
<i>Instructional Management</i> Supplies readily available. Begins lesson without delay. Avoids unnecessary disruptions in lesson. Monitors student time on task. Provides smooth transitions between activities. Paces instruction appropriately. Anticipates classroom disruptions and plans accordingly. Establishes a climate of mutual respect. Anticipates and corrects disruptive behavior in a constructive and timely manner. Recognizes inconsequential behavior and responds accordingly.

Table 3: Beginner teacher performance training videotapes.

<i>Score Points</i>	<i>Frequency of Response</i>	
	<i>First Videotape</i>	<i>Second Videotape</i>
4	0	4
3	1	8
2	8	0
1	3	0
Total	12	12

How many thought it fell in the lower half of the score points?" Then, each rater restated their holistic score and explained how they had arrived at their score. The results were tallied and written on a blackboard for all the raters to see.

The process of viewing training videotapes, scoring, and tally scores continued until the scores clustered together, thus, indicating group consensus. It took the raters two 30 minute beginner teacher performance training videotapes to reach consensus. The results for the two beginner teacher performance training videotapes are presented in Table 3. It can be seen that the raters become more reliable in score assignment by the second beginner teacher performance videotape.

The Holistic Scoring Session Once consensus had been reached, the raters, paired off and went to six different rooms. Each of the rooms contained two 13 inch television monitors and two VCR machines. Additionally, each room contained a tape recorder. In this way, each pair of raters could independently score beginner teacher performance videotapes for all three conditions, that is, front view, rear view, and both views. At the completion of each videotape, each pair of raters briefly discussed their scores. The discussion was audiotaped. They, then moved onto the next videotape without further discussion. Additionally, the raters were asked to provide scores at three different time intervals, that of, ten minutes, twenty minutes, and lastly at the end of the videotape, thirty minutes. It should be restated that the raters scored each time interval independently of each other, and only discussed their ratings after recording their thirty minute score. This was done to determine just how much of a videotape

Table 4: Distribution of the videotapes to rater pairs.

Tape Set One	Rater Pairs									Tape Set Two	Rater Pairs								
	1			2			3				4			5			6		
	A	B	C	A	B	C	A	B	C		A	B	C	A	B	C	A	B	C
1	*				*				*	13	*				*				*
3		*				*	*			15		*			*		*		
4			*	*				*		17			*	*				*	
5	*				*				*	18	*				*				*
6		*				*	*			20		*			*	*			
7			*	*				*		21			*	*			*		
8	*				*				*	22	*				*				*
11		*				*	*			25		*			*	*			
12			*	*				*		24			*	*				*	

needs to be viewed in order for raters to reach consensus, and also to determine just how much of a beginner teacher performance videotape needs to be viewed before raters are confident about their decisions (e.g., their scores).

Raters, then viewed and scored randomized videotapes, independently. Upon completion of viewing and scoring beginner teacher performance videotapes the raters completed a Rater Evaluation Form. Table 4 illustrates the method of distribution of beginner teacher performance videotapes where A is the front camera condition, B the is back camera condition, and C is the dual camera condition. Note that the stars illustrate which camera position the raters pairs viewed for the beginner teacher performance videotapes. Additionally, each number in the table represents a beginner teacher performance videotaping session. Whereas, Table 5 shows the viewing time schedule for each rater pair. Note that the six rater pairs viewed only one videotaping condition at one time, that is, either the front camera condition, the back camera condition, or the dual camera condition for each beginner teacher performance videotape. Also, the numbers in Table 5 represent a beginner teacher performance videotaping session. For example rater pair one at ten thirty viewed a front camera condition beginner teacher performance videotaping session for beginner teacher one.

Table 5: Videotape viewing schedule for rater pairs.

<i>Start Time</i>	<i>Rater Pairs</i>					
	<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>	<i>5</i>	<i>6</i>
10:30	1-Front	1-Back	5-Dual	13-Front	13-Back	18-Dual
11:00	3-Back	6-Dual	3-Front	15-Back	20-Dual	15-Front
11:30	7-Dual	4-Front	12-Back	21-Dual	24-Front	24-Back
12:00	5-Front	5-Back	1-Dual	18-Front	18-Back	13-Dual
14:00	6-Back	3-Dual	6-Front	20-Back	15-Dual	20-Front
14:30	12-Dual	7-Front	7-Back	24-Dual	21-Front	21-Back
15:00	4-Dual	8-Back	11-Front	17-Dual	22-Back	25-Front
15:30	11-Back	12-Front	8-Dual	25-Back	17-Front	22-Dual
16:00	8-Front	11-Dual	4-Back	22-Front	25-Dual	17-Back

3 RESULTS

3.1 The Holistic Scoring Scale

The questions that were raised in reference to the construction of the holistic scoring scale are as follows:

1. Can a holistic scoring scale be developed that accurately reflects beginner teacher performance?
2. Is it possible to train raters to accurately score beginner teacher performance videotapes?
3. How much of a thirty minute beginner teacher performance videotape needs to be viewed by raters in order to establish acceptable levels of inter-rater reliability?

The first step in answering these questions was to analyze the score given to the beginner teacher performance videotapes by the raters. Thus, a Pearson product-moment correlation between rater pairs at three points in time, that of 10 minutes, 20 minutes, and 30 minutes, was performed. The results are presented in Figure 3.1.

Figure 3.1 suggests that a significant relationship exists between the rater pairs at 10 minutes ($r=.91$, $p < .01$), at 20 minutes ($r=.96$, $p < .01$), and at 30 minutes ($r=.99$, $p < .01$). Further, the average scores given by first raters at ten minutes, twenty minutes, and

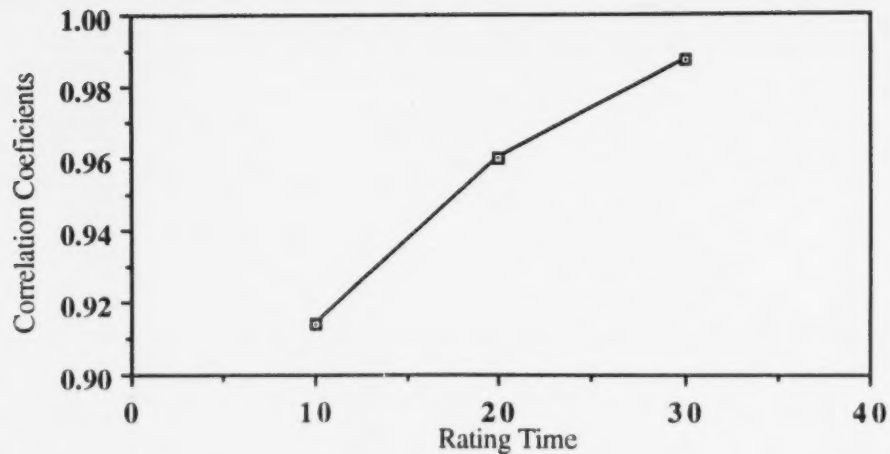


Figure 2: The correlation between rater pairs over time.

thirty minutes were 2.67 (SD=1.08), 2.70 (SD=1.19), and 2.74 (SD=1.20), respectively. Whereas, the average scores given by second raters at ten minutes, twenty minutes, and thirty minutes were 2.69 (SD=1.11), 2.67 (SD= 1.17), and 2.74 (SD=1.22), respectively. Note that first raters were the raters who viewed and scored the videotape initially, and second raters viewed and the scored the videotapes the second time around. It can be seen that the scores on the average tend to increase over time for both rater groups. This suggests that the more of a beginner teacher performance videotape that the raters are able to view the more similar their scores appear to be. Additionally, it can be seen that the scores are very similar between rater groups which helps to explain the high correlation coefficients that were obtained in Figure 3.1. It also suggests that the holistic scale is a reliable measure.

To further examine the scores given by the rater pairs, the researcher calculated the percentage of agreement between the rater pairs across time. The results are presented in Table 6.

Thus in reference to Table 6 , the rater pairs agreement ranges from 67% to 89% with a total of 80% at ten minutes, 78% to 100% with a total of 89% at twenty minutes, and lastly 89% to 100% with a total of 96% at thirty minutes. Thus, the rater pairs holistic scores become more reliable over time.

The researcher, then examined the percentage of agreement between rater pairs by camera condition, that is, the camera in the front of the room condition, the camera in the back of the room condition, and the dual camera condition at 10 minutes, 20 minutes and 30 minutes. The results are presented in Tables 7 - 9. The results from Tables 7 - 9 suggest that no differences exist due to camera condition. However, there appears to be a difference due to time such that the percentage of agreement between

Table 6: Percentage of Agreement by rater pairs over time.

<i>Rater Pairs</i>	<i>10 Minutes</i>	<i>20 Minutes</i>	<i>30 Minutes</i>
1	67	89	89
2	67	100	100
3	89	89	100
4	78	89	100
5	89	78	89
6	89	89	100
Total	80	89	96

rater pairs increases over time.

It should also be noted that the average total percentage of agreement that was found in Tables 7 - 9 is the same totals that were obtained in Table 6.

In sum, the analyses of the data suggest that there are no differences in terms of reliability due to camera condition. However, the analyses suggest that the rater pair scores become more consistent, or more reliable over time.

3.2 Videotaping Conditions

The results obtained in the previous section suggest that there are no differences in the scores assigned by rater pairs due to camera condition. To further examine this finding a repeated measures analyses of variance was performed at ten minutes, twenty minutes and thirty minutes. It should also be noted that due to the high correlation coefficients obtained in Figure 3.1 the rater pair scores were averaged in preparation for the statistical analyses. The repeated measures analyses of variance also helped to answer the following questions and hypotheses:

Questions

1. Where should the video cameras be placed in order to obtain the most information about beginner teacher performance at 10 minutes, 20 minutes, and 30 minutes?
2. What is the most cost effective method of videotaping?

Hypotheses

Table 7: Percentage of Agreement by rater pairs at ten minutes.

<i>Rater Pairs</i>	<i>Camera Conditions</i>		
	<i>Front</i>	<i>Back</i>	<i>Dual</i>
1	67	67	67
2	33	67	100
3	100	100	67
4	100	67	67
5	100	67	100
6	67	100	100
Total	78	78	83

1. It was hypothesized that beginner teacher performance videotapes obtained from the camera in the back of the room condition would be the most reliable videotaping condition. This prediction was based on the view that more information on beginner teacher behavior could be obtained from observing the beginner teacher's front (i.e., face, not back of the head).
2. It was hypothesized that beginner teacher performance videotapes obtained from the camera in the back of the room would not be scored more reliably than the camera in the front of the room condition. This prediction was based on the view that little information on beginner teacher behavior could be obtained from viewing a beginner teacher's back (i.e., back of the head). However, this condition was likely to yield information on beginner teacher- student interaction and for this reason it was incorporated into the study.
3. It was hypothesized that beginner teacher performance videotapes obtained from the dual image condition would not be more reliably scored than the camera in the front of the room condition. This prediction was based on the view that trained raters would find the dual screen condition too busy, that is they would not be able to process all the information in this condition. However, this condition was likely to yield information on beginner teacher- student interactions. It may also provide information on beginner teacher behavior that the other two videotaping conditions independently could not achieve.

Table 8: Percentage of Agreement by rater pairs at twenty minutes.

<i>Rater Pairs</i>	<i>Camera Conditions</i>		
	<i>Front</i>	<i>Back</i>	<i>Dual</i>
1	67	100	100
2	100	100	100
3	100	100	67
4	100	100	67
5	67	67	100
6	100	67	100
Total	89	89	89

The results of the repeated measures analyses of variance are presented in Tables 10 - 12.

The results in Table 10 suggest that there are no differences due to camera condition $F(2, 46)=1.99, p < .15$, or rater pairs $F(5, 46)=.78, p < .56$.

Similarly, the results in Table 11 suggest that there are no differences due to camera condition $F(2, 46)=1.02, p < .37$, or rater pairs $F(5, 46)=.46, p < .80$.

Lastly, the results in Table 12 suggest that there are no differences due to camera condition $F(2, 46)=1.12, p < .34$, or rater pairs $F(5, 46)=.64, p < .67$.

In sum, the constructed repeated measures analyses of variance tables suggest that there are no significant differences between camera conditions or between rater pairs at 10 minutes, 20 minutes, and 30 minutes. However, the raters did have a preferred camera condition such that when asked the following question:

In your opinion, which camera position (a-front, B-back, or C- simultaneous) more accurately recorded beginner teacher performance?

The results are presented in Table 13. It can be seen that three or 25% of the raters felt that having the camera in the front of the room more accurately recorded beginner teacher performance. Another, three or 25% of the raters felt that having the camera in the back of the room more accurately recorded beginner teacher performance. Whereas, six or 50% of the raters felt that having both camera, front and back, more accurately recorded beginner teacher performance. Reasons for not wanting to use both camera positions by the raters, were due to difficulty they found in getting the videotapes to run simultaneously (i.e., at the same time). Additionally, they found that they could

Table 9: Percentage of Agreement by rater pairs at thirty minutes.

<i>Rater Pairs</i>	<i>Camera Conditions</i>		
	<i>Front</i>	<i>Back</i>	<i>Dual</i>
1	100	100	67
2	100	100	100
3	100	100	100
4	100	100	100
5	100	67	100
6	100	100	100
Total	100	94	94

not easily view both television monitors at the same time "it became too busy".

Thus, in reference to the first question:

Where should the video cameras be placed in order to obtain the most information about beginner teacher performance at 10 minutes, 20 minutes, and 30 minutes?

The cameras should be placed at the front and the back of the classroom to be videotaped in, since the majority of the raters (50%) felt that this position more accurately recorded beginner teacher performance. Further, the raters felt this position was best because it was not always possible to obtain enough information from one camera. However, with two cameras there was a greater likelihood of getting the necessary information on beginner teacher performance to reliably score the videotape.

In reference to the second question:

What is the most cost effective method of videotaping?

The statistical analyses of the data suggests that there are no differences in camera conditions. Further, fifty percent of the raters prefer to use only one camera as opposed to 50% of the raters, who prefer two use the dual camera condition. Thus, it stands to reason that the most economical method of videotaping would be to use only one camera, since half the raters prefer to view beginner teacher performance videotapes from one camera position, and additionally, since there are no statistical differences between the camera conditions or between the rater pairs.

In reference to the three hypotheses, there are no statistically significant differences between the three videotaping conditions. However, the raters did have a preferred viewing condition as illustrated in Table 13.

Table 10: Repeated measures analysis of variance at ten minutes.

<i>Source of Variation</i>	<i>SS</i>	<i>DF</i>	<i>MS</i>	<i>F</i>	<i>Sig.</i>
Main Effects	8.977	7	1.282	1.132	0.360
Camera	4.509	2	2.255	1.991	0.148
Raters	4.468	5	0.894	0.789	0.563
Explained	8.977	7	1.282	1.132	0.360
Residual	52.102	46	1.133		
Total	61.079	53	1.152		

In closing, there were no statistically significant differences between camera conditions, or between rater pairs at 10 minutes, 20 minutes, and at 30 minutes. However the raters, themselves, suggested a preference for a particular camera condition as was illustrated in Table 13. These results as a whole suggest that when time and money are not a problem, then two cameras should be used. However, when time and money are in short supply, then one camera is called for. Additional savings can be obtained when the camera(s) utilize a wide-lens and are mounted on tripods and left to run independently.

4 DISCUSSION OF RESULTS

4.0.1 Conclusion

Thus, it appears that teacher performance videotapes can be reliably scored by trained raters using a four point holistic scoring scale. The results also suggest that a twenty to thirty minute videotape provides the raters with ample teacher performance information for obtaining consistent judgments by raters.

Further, the results suggest that there are no differences due to camera condition. However, raters do have a preferred viewing condition. For this reason when time and money are not a major concern then two cameras can be used. However, one camera is more cost effective and is capable of providing sufficient teacher performance information for obtaining reliable scores.

Table 11: Repeated measures analysis of variance at twenty minutes.

<i>Source of Variation</i>	<i>SS</i>	<i>DF</i>	<i>MS</i>	<i>F</i>	<i>Sig.</i>
Main Effects	6.241	7	0.892	0.622	0.735
Camera	2.926	2	1.463	1.021	0.368
Raters	3.315	5	0.663	0.463	0.802
Explained	6.241	7	0.892	0.622	0.735
Residual	65.907	46	1.433		
Total	72.148	53	1.361		

4.1 Validity Issues in Holistic Scoring

Validity deals with the truthfulness of a particular test or measure, a valid measure demonstrates a connection between what it purports to be testing and what it is actually testing (Williamson, 1993). The Standards for Educational and Psychological Testing (1985) states, "Validity is the most important consideration in test evaluation. The concept refers to the appropriateness, meaningfulness, and usefulness of the specific inferences made from the test scores (*p.9*)."

Further, Anastasi (1988) suggests that there are three main types of validity such that:

Content-related validation involves essentially the systematic examination of the test content to determine whether it covers a representative sample of the behavior domain to be measured.

Criterion-related validation procedures indicate the effectiveness of a test in predicting an individual's performance in specified activities.

Construct-related validation of a test is the extent to which the test may be said to measure a theoretical construct or trait.

In terms of this study, the holistic scoring scale may be said to have a measure of content validity, since the researcher based the scale on current teaching literature, Analytic Scales, the beginner teacher lesson plans, and the collected beginner teacher videotapes. It can also be said to have face validity. Anastasi (1988) writes that this not validity in its strictest sense, since "it refers, not to what the test actually measures,

Table 12: Repeated measures analysis of variance at thirty minutes.

<i>Source of Variation</i>	<i>SS</i>	<i>DF</i>	<i>MS</i>	<i>F</i>	<i>Sig.</i>
Main Effects	8.102	7	1.157	0.774	0.612
Camera	3.343	2	1.671	1.118	0.336
Raters	4.759	5	0.952	0.637	0.673
Explained	8.102	7	1.157	0.774	0.612
Residual	68.769	46	1.495		
Total	76.870	53	1.450		

Table 13: Raters' preferred camera condition.

<i>Camera Condition</i>	<i>Number of Raters</i>	<i>Percentage of Raters</i>
Front	3	25
Back	3	25
Dual	6	50

but what it appears superficially to measure (*p.* 44)." Nevertheless, it is desirable to have a test that looks like what it is supposed to be given that no form of evaluation is context free. White (1983) notes that writing assessment puts a lot of weight in the value of face validity.

This study does not have criterion-related validity given the fact that no additional data was collected at the time of the study (concurrent validity) such as grade point average. Nor was any data collected from the beginner teachers after the initial collection of the videotapes (predictive validity) such as on the job performance measures.

Lastly, this study does not have construct validity because as of today, there exists no unifying body of information related to teacher performance, in other words, there is no single theory of teaching per se. This is also true of writing assessment. Even so, Williamson (1993) suggests that this is the most important type of validity. However, this study and writing assessment utilizing holistic scoring did make use of the literature

that was available from their perspective fields as Williamson (1993) states:

The immediate need for writing assessment demanded the development of assessments constructed within the limitations of current theory and within the fiscal limitations imposed by specific educational institutions. At that, there was little time, for further study and only time for standing one's ground. . . (p.14).

In sum, researchers and test developers alike must make use of what knowledge in their perspective fields is currently available.

It should also be noted that some people may protest that being videotaped while teaching a lesson is an artificial type of performance assessment. It must be agreed that this type of testing situation or any other kind of testing condition does indeed represent a limited form of reality. However, this does not mean that holistic scoring of beginner teacher performance videotapes is not a valid form of assessment, only that like any other test it has limited value such that it only elludes to the potential of a beginner teacher. Although, it may be argued that teaching before a camera is not real teaching, it may be said to be no worse if not better than direct observation given that people tend to become desentized to having a camera in the room. However, this is not true of having an outside observer in the room. Additionally, holistic scoring of beginner teacher performance videotapes is measuring a product (the one lesson on videotape), not the teaching process.

4.2 Recommendations

The beginner teachers in this study were asked only to submit a videotape of a lesson they taught along with a lesson plan for that lesson. They were not given the opportunity to try out various lessons on videotape, nor were they given much information about how the videotape would be used in the study.

However, if this were a large scale assessment test it would be necessary for the perspective examinees to know and understand the four point holistic scoring scale in order to be fully aware of how their teaching performance videotape would be judged by the raters. They should also be given ample time and practice to obtain the skills that the test is assessing through inservice teacher workshops and review of printed material discussing the teacher performance assessment. This is where self- and peer videotaping could serve as a model for perspective examinees to practice the skills that the test is assessing as illustrated by the description of score points. Further, an abbreviated form of microteaching, that is, a form of microteaching (Primary Trait Method) that does not

impose severe time restrictions but does call for a more focused type of assessment (e.g., how to use guided practice in a lesson) may be useful in allowing teachers to practice the skills that make up the objectives and the description of score points.

Lastly, and most importantly, perspective examinees should be able to view videotaped exemplars that illustrate each of the score points, so that they can become familiar with what is being assessed. They should also be given the opportunity to videotape as many of their own lessons as they wish and to submit the one videotape they feel best represents their teaching performance. This is due to the fact that all assessment measures are looking for optimum performance. However, even if a perspective examinee is able to perform optimally that does not preclude that she/he always teaches in this manner, only that she/he can or has the skills necessary to teach.

4.3 Future Research

This is a relatively new research area and therefore the number of studies that could be conducted are numerous. In general, though, there needs to be research on the validity and reliability of holistic scoring for both beginner and experienced teachers in order to ascertain just what the parameters are in this area of study.

There should also be studies dealing with the issue of holistic versus analytic scoring of performance videotapes.

Research on direct observation versus videotaped observation utilizing holistic scoring.

There needs to be research on who should be the raters and is it necessary for raters to be familiar with the content presented in the videotape.

Additionally, there needs to be research in how raters arrive at their score or rating. Specifically, what is the process that raters go through before assigning a score in holistic scoring.

Holistic scoring does not provide diagnostic information on a teacher's performance. However, at times this is just what is needed (e.g., when an examinee fails an exam). Thus, future research should explore the primary trait method which is a more focused or refined type of holistic scoring to ascertain if it is capable of providing diagnostic information on an examinee's performance.

References

- American Psychological Association (1985). *Standards for educational and psychological testing*.
- Anastasi, A. (1988). *Psychological Testing* (6th ed.). New York: Macmillan Publishing Company.
- Berk (1986). Definition of performance assessment. In Berk (Ed.) *Performance assessment: Method and applications*. John Hopkins University Press.
- Berliner, D. and Carter, K. (1989). Differences in processing classroom information by expert and novice teachers. In J. Lowyck and C. Clark (Eds.). *Teacher thinking and professional action*. Leuven, Belgium: Leuven University Press.
- Biehler, R. and Snowman, J. (1986). *Psychology applied to teaching*. (5th ed.). Boston, Massachusetts: Houghton Mifflin Company.
- Borko, H. and Livingston, C. (1989). Cognition and improvisation: Differences in mathematics instruction by expert and novice teachers. *American Educational Research Journal*, 26, 473-498.
- Borko, H., Livingston, C., McCaleb, J., and Mauro, L. (1988). Student teachers' planning and post-lesson reflections: Patterns and implications for teacher preparation. In J. Calderhead (Ed.). *Teachers' professional learning* (pp. 55-83). London: Falmer.
- Brophy, J., Good, T. (1986). Teacher behavior and student achievement. In M. C. Wittrock (Ed.). *Handbook of research on teaching* (3rd ed., pp.328-375). New York: Macmillan Publishers.
- Chase, W. and Simon, H. (1973). The mind's eye in chess. In W. Chase (Ed.). *Visual information processing*. New York: Academic Press.
- Cherry, R. and Meyer, P. (1993). Reliability issues in holistic assessment. In M. Williamson and B. Hout (Eds.), *Validating holistic scoring for writing assessment. Theoretical and empirical foundations*. (pp. 1-43). NJ: Hampton, Press, Inc.
- Chi, M., Feltovich, A. and Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science*, 5, 121-152.
- Clark, C. and Peterson, P. (1986). Teachers' thought processes. In M. C. Wittrock (ed.). *Handbook of research on teaching* (3rd ed., pp. 255-296). New York: Macmillan Publishers.
- Commonwealth of Massachusetts Department of Education (1994). *Recertification guide for Massachusetts educators*. (Publication No.: 17596-5-150000-8/94,113-DOE).
- Copeland, W. (1981). Clinical experience in the education of teachers'. *Journal of Education for Teaching*, 7(1).
- De Groot, A. (1965). *Thought and choice in chess*. The Hague: Mouton.
- Duke, D. and Meckel, A. (1984). *Teacher's guide to classroom management*. New York: Random House Incorporated.

Kounin, J. (1970). *Discipline and group management in classrooms*. New York: Holt, Rinehart, and Winston.

Millman, Jason (1981). *Handbook of Teacher Evaluation*.

Mullis, I. V. S. (1980). *Using the primary trait system for evaluating writing. National Assessment of Educational Progress Report*. Denver, CO: Education Commission of the States.

Needels, M. (1991). Comparison of student, first-year, and experienced teachers' interpretations of a first-grade lesson. *Teaching & Teacher Education*, 7(3), 269-278.

Peters, E. (1979). *Holistic scoring manual. Basic skills improvement policy implementation guide # 2*. Massachusetts State Department of Education, Boston. Bureau of Research and Assessment. (ERIC Document Reproduction Service No. ED 189131).

Reynolds, A. (1992). What is competent beginning teaching? A review of the literature. *Review of Educational Research*, 62(1), 1-35.

Sabers, D., Cushing, K., and Berliner, D. (1991). Differences among teachers in a task characterized by simultaneity, multidimensionality, and immediacy. *American Educational Research Journal*, 28(1), 63-88.

Sandefur, J. (1987). Historical perspective. In L. Rudner, What's happening in teacher testing an analysis of state teacher testing practices. Superintendent of Documents, U.S. Government Printing Office, Washington, DC. (ERIC Document Reproduction Service No. ED 284867).

Schram, P., Feiman-Nemser, S. and Ball, D. (1989). *Think about teaching subtraction with regrouping: A comparison of beginning and experienced teachers' responses to textbooks* (Research Rep. No. 89-5). East Lansing, Michigan: National Center for Research on Teacher Education.

Shulman, J. (1989). Blue freeways: Traveling the alternate route with big-city teacher trainees. *Journal of Teacher Education*, 40, 2-8.

Shulman, L. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review*, 57(1), 1-22.

Siegler, R. (1991). *Children's thinking*. (2nd ed.). New Jersey: Prentice Hall.

Skinner, B. (1968). *The Technology of Teaching*. New York: Appleton, Century, and Crofts.

SPSS (1990). *Base system user's guide*. SPSS, Inc.

Stanley, S. (1990). Videotaped-based teacher tests: Bells, whistles or bona fide assessment? Paper presented at the Annual Meeting of the American Education Research Association, Boston, MA. (ERIC Document Reproduction Service No. ED 319805).

State Education Department/University of the State of New York (1991, July 19). State Education Department awards contract for new teacher certification exam (News release). Albany, NY: Office of Communications.